

BlackboxNLP2020

**Proceedings of the Third BlackboxNLP Workshop on
Analyzing and Interpreting Neural Networks for NLP**

©2020 The Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-952148-86-6

Introduction

BlackboxNLP is the workshop on analyzing and interpreting neural networks for NLP.

In the last few years, neural networks have rapidly become a central component in NLP systems. The improvement in accuracy and performance brought by the introduction of neural networks has typically come at the cost of our understanding of the system: How do we assess what the representations and computations are that the network learns? The goal of this workshop is to bring together people who are attempting to peek inside the neural network black box, taking inspiration from machine learning, psychology, linguistics, and neuroscience.

In this third edition of the workshop, hosted by the 2020 conference on Empirical Methods in Natural Language Processing (EMNLP), we accepted 31 archival papers and 9 extended abstracts. The workshop also provided a platform for authors of EMNLP-Findings papers to present their work as a poster at the workshop. Lastly, for the first time, BlackboxNLP included a shared interpretation mission. One paper submitted to this mission has a *demo* presentation, of the interpretability library `diagnose`, and is included as the last paper in these proceedings (submission number 70).

BlackboxNLP would not have been possible without the dedication of its program committee. We would like to thank them for their invaluable effort in providing timely and high-quality reviews on a short notice. We are also grateful to our invited speakers for contributing to our program.

Afra Alishahi, Yonatan Belinkov, Grzegorz Chrupała, Dieuwke Hupkes, Yuval Pinter and Hassan Sajjad

Organizers:

Afra Alishahi, Tilburg University
Yonatan Belinkov, Technion - Israel Institute of Technology
Grzegorz Chrupała, Tilburg University
Dieuwke Hupkes, University of Amsterdam
Yuval Pinter, Georgia Institute of Technology
Hassan Sajjad, Qatar Computing Research Institute

Program Committee:

Samira Abnar, University of Amsterdam
Željko Agić, Unity Technologies Copenhagen
Antonios Anastasopoulos
Leila Arras - Fraunhofer Heinrich Hertz Institute
Jasmijn Bastings - Google
Lisa Beinborn - University of Amsterdam
Laurent Besacier - Laboratoire d'Informatique de Grenoble
Stergios Chatzikyriakidis - University of Gotheburg
Barry Devereux - Queen's University
Ewan Dunbar - Université Paris Diderot
Allyson Ettinger - University of Chicago
Antske Fokkens - Vrije Universiteit Amsterdam
Robert Frank - Yale University
Alexander Fraser - LMU Munich
Richard Futrell - University of California, Irvine
Sebastian Gehrmann - Harvard University
David Harwath - MIT
John Hewitt - Stanford University
Cassandra Jacobs - University of Wisconsin
Yair Lakretz - NeuroSpin
Shalom Lappin - University of Gothenburg
Miryam de Lhoneux - Uppsala University
Tal Linzen - Johns Hopkins University
Nelson F. Liu - University of Washington
Pranava Madhyastha - Imperial College London
Arya McCarthy - Johns Hopkins University
Paola Merlo - University of Geneva
Raymond Mooney - UT Austin Joakim Nivre - Uppsala University
Sebastian Padó - University of Stuttgart
Ellie Pavlick - Brown University
Rudolf Rosa - Charles University
Carolyn Rose - CMU Sebastian Ruder - DeepMind
Wojciech Samek - Fraunhofer Heinrich Hertz Institute
Naomi Saphra - University of Edinburgh
Sabine Schulte im Walde - University of Stuttgart
Rico Sennrich - University of Zurich
Pia Sommerauer - Vrije Universiteit Amsterdam

Ivan Titov - University of Edinburgh
Francesca Toni - Imperial College London
Reut Tsarfaty - Open University
Sarah Wiegrefe - Georgia Tech
Adina Williams - New York University
Diyi Yang - Georgia Tech
Fabio Massimo Zanzotto - University of Rome

Invited Speakers:

Idan Blank, UCLA
Roger Levy, MIT
Anna Rogers, University of Copenhagen

Table of Contents

| | |
|---|-----|
| <i>BERTering RAMS: What and How Much does BERT Already Know About Event Arguments? - A Study on the RAMS Dataset</i> | |
| Varun Gangal and Eduard Hovy | 1 |
| <i>Emergent Language Generalization and Acquisition Speed are not tied to Compositionality</i> | |
| Eugene Kharitonov and Marco Baroni | 11 |
| <i>Examining the rhetorical capacities of neural language models</i> | |
| Zining Zhu, Chuer Pan, Mohamed Abdalla and Frank Rudzicz | 16 |
| <i>What Happens To BERT Embeddings During Fine-tuning?</i> | |
| Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick and Ian Tenney | 33 |
| <i>It's not Greek to mBERT: Inducing Word-Level Translations from Multilingual BERT</i> | |
| Hila Gonen, Shauli Ravfogel, Yanai Elazar and Yoav Goldberg | 45 |
| <i>Leveraging Extracted Model Adversaries for Improved Black Box Attacks</i> | |
| Naveen Jafer Nizar and Ari Kobren | 57 |
| <i>On the Interplay Between Fine-tuning and Sentence-Level Probing for Linguistic Knowledge in Pre-Trained Transformers</i> | |
| Marius Mosbach, Anna Khokhlova, Michael A. Hedderich and Dietrich Klakow | 68 |
| <i>Unsupervised Evaluation for Question Answering with Transformers</i> | |
| Lukas Muttenthaler, Isabelle Augenstein and Johannes Bjerva | 83 |
| <i>Unsupervised Distillation of Syntactic Information from Contextualized Word Representations</i> | |
| Shauli Ravfogel, Yanai Elazar, Jacob Goldberger and Yoav Goldberg | 91 |
| <i>The Explanation Game: Towards Prediction Explainability through Sparse Communication</i> | |
| Marcos Treviso and André F. T. Martins | 107 |
| <i>Latent Tree Learning with Ordered Neurons: What Parses Does It Produce?</i> | |
| Yian Zhang | 119 |
| <i>Linguistically-Informed Transformations (LIT): A Method for Automatically Generating Contrast Sets</i> | |
| Chuanrong Li, Lin Shengshuo, Zeyu Liu, Xinyi Wu, Xuhui Zhou and Shane Steinert-Threlkeld | 126 |
| <i>Controlling the Imprint of Passivization and Negation in Contextualized Representations</i> | |
| Hande Celikkanat, Sami Virpioja, Jörg Tiedemann and Marianna Apidianaki | 136 |
| <i>The elephant in the interpretability room: Why use attention as explanation when we have saliency methods?</i> | |
| Jasmijn Bastings and Katja Filippova | 149 |
| <i>How does BERT capture semantics? A closer look at polysemous words</i> | |
| David Yenicelik, Florian Schmidt and Yannic Kilcher | 156 |
| <i>Neural Natural Language Inference Models Partially Embed Theories of Lexical Entailment and Negation</i> | |
| Atticus Geiger, Kyle Richardson and Christopher Potts | 163 |

| | |
|---|-----|
| <i>BERTnesia: Investigating the capture and forgetting of knowledge in BERT</i> Jaspreet Singh, Jonas Wallat and Avishek Anand | 174 |
| <i>Probing for Multilingual Numerical Understanding in Transformer-Based Language Models</i> Devin Johnson, Denise Mak, Andrew Barker and Lexi Loessberg-Zahl | 184 |
| <i>Dissecting Lottery Ticket Transformers: Structural and Behavioral Study of Sparse Neural Machine Translation</i> Rajiv Movva and Jason Zhao | 193 |
| <i>Exploring Neural Entity Representations for Semantic Information</i> Andrew Runge and Eduard Hovy | 204 |
| <i>BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance</i> R. Thomas Mccoy, Junghyun Min and Tal Linzen | 217 |
| <i>Attacking Semantic Similarity: Generating Second-Order NLP Adversarial Examples</i> John Morris | 228 |
| <i>Discovering the Compositional Structure of Vector Representations with Role Learning Networks</i> Paul Soulos, R. Thomas Mccoy, Tal Linzen and Paul Smolensky | 238 |
| <i>Structured Self-Attention Weights Encodes Semantics in Sentiment Analysis</i> Zhengxuan Wu, Thanh-Son Nguyen and Desmond Ong | 255 |
| <i>Investigating Novel Verb Learning in BERT: Selectional Preference Classes and Alternation-Based Syntactic Generalization</i> Tristan Thrush, Ethan Wilcox and Roger Levy | 265 |
| <i>The EOS Decision and Length Extrapolation</i> Benjamin Newman, John Hewitt, Percy Liang and Christopher D. Manning | 276 |
| <i>Do Language Embeddings capture Scales?</i> Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar and Dan Roth | 292 |
| <i>Evaluating Attribution Methods using White-Box LSTMs</i> Yiding Hao | 300 |
| <i>Defining Explanation in an AI Context</i> Tejaswani Verma, Christoph Lingensfelder and Dietrich Klakow | 314 |
| <i>Searching for a Search Method: Benchmarking Search Algorithms for Generating NLP Adversarial Examples</i> Jin Yong Yoo, John Morris, Eli Lifland and Yanjun Qi | 323 |
| <i>This is a BERT. Now there are several of them. Can they generalize to novel words?</i> Coleman Haley | 333 |
| <i>diagNNose: A Library for Neural Activation Analysis</i> Jaap Jumelet | 342 |

